Published in final edited form as:

Nat Genet. 2016 October ; 48(10): 1193–1203. doi:10.1038/ng.3646.

Lineage-specific and single cell chromatin accessibility charts human hematopoiesis and leukemia evolution

M. Ryan Corces[1,2,3,10], **Jason D. Buenrostro**[3,4,10,11], **Beijing Wu**[4], **Peyton G. Greenside**[4,5], **Steven M. Chan**[6], **Julie L. Koenig**[1,2], **Michael P. Snyder**[3,4], **Jonathan K. Pritchard**[4,7,8], **Anshul Kundaje**[4,9], **William J. Greenleaf**[3,4], **Ravindra Majeti**[1,2,11], and **Howard Y. Chang**[3,11]

[1]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA

[2]Department of Medicine, Division of Hematology, Stanford University School of Medicine, Stanford, CA 94305, USA

[3]Center for Personal Dynamic Regulomes, Stanford University School of Medicine, Stanford, CA 94305, USA

[4]Department of Genetics, Stanford University, Stanford, CA 94305, USA

[5]Program in Biomedical Informatics, Stanford University School of Medicine, Stanford, CA 94305, USA

[6]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada

[7]Department of Biology, Stanford University, Stanford, CA 94305, USA

Correspondence should be addressed to: R.M. (rmajeti@stanford.edu) or H.Y.C. (howchang@stanford.edu).

[10]These authors contributed equally to this work

[11]These authors jointly directed this work

Contact Information: Ravindra Majeti MD, PhD, Stanford University School of Medicine, Stanford Institute for Stem Cell Biology and Regenerative Medicine, Lokey Stem Cell Building, 265 Campus Drive, Stanford, CA 94305-5463, rmajeti@stanford.edu, Phone: 650-721-6376, Fax: 650-736-2961.

Howard Y. Chang, MD, PhD, Stanford University School of Medicine, CCSR 2155c, 269 Campus Drive, Stanford, CA 94305-5168, howchang@stanford.edu, Phone: 650-736-0306, Fax: 650-723-8762

URLs

Jaspar Website - http://jaspar.genereg.net/

UCSC Genome Browser Track Hub URL - https://s3-us-west-1.amazonaws.com/chang-public-data/2016_NatGen_ATAC-AML/hub.txt

ACCESSION CODES

All ensemble ATAC- and RNA-seq data is available through GEO accession number GSE74912. We provide raw sequencing reads, processed BAM files, and fully processed count matrices for ATAC-seq and RNA-seq at this accession. All single cell ATAC-seq data is available through GEO accession number GSE74310. All analyses and coordinates referenced here are for the human reference genome hg19.

AUTHOR CONTRIBUTIONS

M.R.C., J.D.B., R.M., H.Y.C. conceived the project. M.R.C. performed all cell sorting, RNA-seq, CIBERSORT analysis, AML cell culture experiments, and mouse experiments. J.D.B. performed all ATAC-seq data analysis and regulatory network analysis and oversaw all ATAC-seq library generation and protocol optimization performed by B.W. M.R.C. and J.L.K. performed DNA genotyping for AML patients. J.D.B., P.G.G., A.K. performed GWAS correlation analyses. W.J.G., M.P.S., J.K.P. assisted with sequencing and study design. S.M.C. collected patient follow-up data and performed all survival analyses. M.R.C., J.D.B., R.M., and H.Y.C. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

[8]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

[9]Department of Computer Science, Stanford University, Stanford, CA 94305, USA

Abstract

We define the chromatin accessibility and transcriptional landscapes in thirteen human primary blood cell types that traverse the hematopoietic hierarchy. Exploiting the finding that the enhancer landscape better reflects cell identity than mRNA levels, we enable “enhancer cytometry” for enumeration of pure cell types from complex populations. We identify regulators governing hematopoietic differentiation and further reveal the lineage ontogeny of genetic elements linked to diverse human diseases. In acute myeloid leukemia (AML), chromatin accessibility reveals unique regulatory evolution in cancer cells with progressive mutation burden. Single AML cells exhibit distinctive mixed regulome profiles of disparate developmental stages. A method to account for this regulatory heterogeneity identified cancer-specific deviations and implicated HOX factors as key regulators of pre-leukemic HSC characteristics. Thus, regulome dynamics can provide diverse insights into hematopoietic development and disease.

INTRODUCTION

The entire human hematopoietic system is maintained by a small number of self-renewing multipotent hematopoietic stem cells (HSCs). More than 200 billion blood cells are produced in a single day[1], highlighting the need for exquisite regulation that balances self-renewal of upstream stem cells with downstream production of differentiated effector cells. Previous studies have profiled gene expression patterns in mouse[2,3] and human[4,5] hematopoiesis providing a rich resource for characterizing these cellular states. However, measuring gene expression alone provides limited information regarding the causative regulators of cell identity. Alternatively, genome-wide chromatin-based assays are sensitive methods for assaying the activity of *trans* factors and *cis* regulatory elements. Recently, several methods have been developed to profile the epigenomes of rare cellular populations[3,6,7], enabling the identification of regulatory elements within mouse hematopoiesis[3]. These methods have not yet been used to profile the epigenomes within rare progenitor populations of human hematopoiesis.

Dysregulation of the regulatory networks governing the human hematopoietic system plays a critical role in the development of hematologic malignancies[8]. The long lifespan of HSCs makes them susceptible to the accumulation of mutations over time[9,10]. In particular, in the case of acute myeloid leukemia (AML), HSCs isolated from leukemia patients have been shown to harbor some but not all of the genetic alterations found in leukemic cells. These cells, termed pre-leukemic HSCs[11–13], provide insight into the earliest stages of the dysregulation of normal hematopoiesis leading to AML.

We previously described the Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq), a method capable of measuring chromatin accessibility in rare cellular populations[6]. Here, we report the development of an improved ATAC-seq protocol, optimized for human blood cells, that allows for more rapid high-quality measurements. We apply this optimized protocol to cells isolated from 9 healthy human donors and 12 AML

patients, studying a total of 137 samples representing 16 of the major cell types of the normal hematopoietic and leukemic hierarchies. In addition, we measure the transcriptomes of 96 samples from the same healthy and leukemic donors to derive paired expression data. This reference map revealed the effects of both early mutations in epigenetic modifiers and late mutations in proliferative oncogenes on the leukemogenic process. Our results provide key insights into the evolutionary process of leukemogenesis and identify important regulatory programs that could be targeted to disrupt this process during its earliest stages.

RESULTS

Fast-ATAC is an optimized ATAC-seq protocol for blood cells

We created a reference regulome and transcriptome map of the normal hematopoietic hierarchy (Fig. 1a,b). We developed an optimized protocol for use on primary blood cells, termed Fast-ATAC, which relies on a 1-step membrane permeabilization and transposition using the lysis reagent digitonin. We found that this simplified protocol requires just 5,000 cells, provides high quality data with reduced signal noise (Supplementary Fig. 1a–c), reduces the frequency of mitochondrial reads by ~5 fold (Supplementary Fig. 1d), and offers an approximately 5 fold improvement in fragment yield per cell (Supplementary Fig. 1e).

Using Fast-ATAC and RNA-seq, we profiled the chromatin accessibility landscape (“regulomes”) and transcriptomes from 13 distinct cellular populations from the human hematopoietic hierarchy isolated via fluorescence activated cell sorting (FACS) (Fig. 1a and Supplementary Fig. 2–4). Cells were taken directly from donor bone marrow or peripheral blood without further manipulation (Supplementary Table 1). The isolated cell populations included 7 unique stem and progenitor and 6 differentiated cell types spanning the myeloid, erythroid, and lymphoid lineages[14–17]. All together, we performed ATAC-seq and RNA-seq on 3–4 adult donors for each cell population totaling 49 transcriptomes and 77 regulomes (Fig. 1c, Supplementary Fig. 1f, Supplementary Fig. 5a,b, and Supplementary Table 1).

With this dataset we identified a total of 590,650 accessible peaks. We found Fast-ATAC profiles to be highly reproducible across technical (R=0.98, Fig. 1d) and biological (R=0.97, Fig 1e) replicates within hematopoietic stem cells (HSCs). In addition, we found similarly high concordance across all other cell types for all technical and biological replicates (mean R=0.94 and R=0.91 respectively, Supplementary Fig. 1g,h) except for erythroblast cells (technical replicates, R=0.55; biological replicates, R=0.50). Each individual cell type of the hematopoietic hierarchy displayed a set of uniquely expressed genes and uniquely accessible peaks mapping to genes known to be involved in cellular functions important for the given cell type (Fig. 1c and Supplementary Fig. 6a–c).

We also observed reasonable correlation (R=0.73) between Fast-ATAC and DNase-seq[18] of CD34+ HSPCs (Fig. 1f). Importantly, we find that HSCs, a CD34+ subpopulation, can have different ATAC-seq profiles than the bulk CD34+ HSPC pool (R=0.77 observed versus R=0.91 expected for same cell type replicates, Fig. 1g), highlighting the value of highly purified stem and progenitor cell subpopulations for epigenomic analysis.

Distal element accessibility is highly cell type specific

Unsupervised hierarchical clustering of our RNA-seq and ATAC-seq data shows robust classification of cell types among technical and biological replicates (Fig. 2a–d, Supplementary Fig. 7a–d). In this analysis, we observe chromatin accessibility is more adept than mRNA expression levels at classifying cell types, quantified by cluster purity[19], suggesting that chromatin accessibility is more cell type-specific and better captures cell identity. However, we note that RNA information from enhancer transcription, splicing, or other features that require optimized methods, and deeper sequencing may improve cell type classification. When regulatory elements were subdivided as gene promoters or distal elements (>1,000 bp away from a transcription start site (TSS)), we find that distal elements provide significantly improved cell-type classification compared to promoters (Fig. 2e,f), similar to previous observations using DNase-seq and ChIP-seq data[20,21]. This observation is clearly illustrated by the region surrounding the *TET2* gene. Despite the invariant expression of *TET2* and ubiquitous accessibility of *TET2* promoter, we find highly diverse accessibility profiles within nearby distal regulatory elements, clearly distinguishing HSPCs, NK cells, and T cells (Fig. 2g).

Enhancer cytometry deconvolutes complex cell populations

Given the accuracy with which distal regulatory landscapes delineate cell types, we hypothesized that Fast-ATAC data can be used to deconvolve highly complex cellular populations, such as CD34+ HSPCs, into their constitutive subsets (Fig. 3a). The highly cell type-specific nature of our ATAC-seq data enabled the development of a strategy we term “enhancer cytometry”, wherein we enumerate the frequency of cell types in complex cellular mixtures *in silico* based on chromatin accessibility data. To do this, we employ the deconvolution algorithm CIBERSORT[22] to quantify the contribution of each individual cell type to the ensemble profile (see methods). Using a filtered peak list, we applied CIBERSORT to define a set of cell-type specific regulatory elements (Fig. 3b and Supplementary Table 2). We validated this approach using leave-one-out cross validation and found enhancer cytometry was able to classify all normal hematopoietic cell types (Fig. 3c,d and Supplementary Fig. 8a–g). One exception is the discrimination of HSC and MPP, which share similar epigenomic profiles and therefore showed reasonable but lower accuracy than other cell types (Supplementary Fig. 8a,g). Comparison of enhancer cytometry on bulk CD34+ HSPCs to ground truth flow cytometry data showed accurate enumeration of the constituent cell types (R^2=0.95, Fig. 3e,f). Notably, cell type deconvolution of CD34+ HSPCs using all regulatory elements, including promoters, was not as accurate (R^2=0.91, Supplementary Fig. 8h). In addition, we found that enhancer cytometry can also be used to deconvolve CD34+ DNase-seq data (Supplementary Fig. 8i), suggesting that ATAC-seq with enhancer cytometry may be a general strategy for identifying and enumerating cell types within existing epigenomics data from complex cellular mixtures.

Regulatory networks of normal hematopoiesis

To better understand the mechanisms governing these diverse regulatory landscapes, we sought to quantify the effect of specific trans-factors at each developmental transition. We adapted a computational framework to measure gain or loss of accessibility across

regulatory elements sharing a feature or annotation, for example a transcription factor (TF) motif (see methods)[23]. For subsequent visualization, we cluster similar motifs to create a non-redundant list we call "hematopoiesis TF motifs" (Fig. 4a, N=46; see methods). We find TF motifs such as "GATA", "RUNX", and "SPI1" to be dominant regulators of chromatin accessibility, consistent with published results[24–26] (Fig. 4a and Supplementary Fig. 9a). We find that activation of these TFs is cell-type specific, often displaying step-wise gains across developmental lineages (Supplementary Table 3). This is exemplified by the "GATA" and "PAX" motifs, which are strongly enriched in erythroid and lymphoid lineages respectively (Fig. 4b,c). To validate this approach for determining global TF motif regulators of cell identity, we compared GATA TF footprints[27] between MEPs (GATA high) and common lymphoid progenitors (CLPs) (GATA low) and found that CLPs had no detectable binding at GATA sites when compared to MEPs (Fig. 4d).

We next reasoned that the accessibility of a given TF motif should correlate with the expression of the associated transcription factor throughout hematopoiesis. However, the underlying motif sequence does not identify the precise causative regulator of accessibility at those motif instances. This is a common issue in epigenomic studies and particularly important for cases in which many factors share identical or near-identical TF motifs. To assign motifs to transcription factors, we integrated our ATAC-seq and RNA-seq data to predict causative regulators of motif accessibility (Supplementary Fig. 9b–e and Supplementary Table 4; see methods). Using this approach we find a striking correlation of motif usage with the expression of known master regulators of hematopoiesis (Fig. 4e). For example, the expression of GATA1 and PAX5 are highly correlated with accessibility at GATA and PAX motifs, respectively ($R=0.75$, $P=10^{-18}$ and $R=0.88$, $P=10^{-230}$, Fig. 4e–g and Supplementary Fig. 9f). Interestingly, for some motifs, such as the HOX motif, we find many putative regulators with weak correlations (N=11; Supplementary Fig. 9g,h), suggesting that regulation of HOX accessibility is more complex. We provide the complete list of non-redundant TF deviations, TF motif to gene association table, and gene correlation analysis as an associated resource (Supplementary Table 3, 4 and Supplementary Data).

Regulome profiles chart the ontogeny of human diseases

In addition to enhancing our understanding of developmental gene regulation, the hematopoietic regulome can trace the ontogeny of activity in the noncoding genome that impacts human disease. Many genome-wide association studies (GWAS) have linked diseases to polymorphisms, but have not been able to pinpoint the cells responsible for those phenotypes. By measuring the activity of regulatory elements that overlap regions with predicted sites of functional variation from GWAS, it is now possible to more accurately predict the specific cell types impacted by genetic variants linked to diverse human diseases (Supplementary Fig. 10a–c; see methods and Supplementary Note 1)[28–30]. As an example, polymorphisms linked to mean corpuscular volume (MCV), a measure of the average volume of an erythrocyte cell, are most strongly enriched in erythroblasts (Fig. 4h). Intriguingly, many regions associated with MCV polymorphisms first become accessible at the CMP and MEP stages suggesting that these polymorphisms may exert their effects prior to full erythroid lineage commitment. Similarly, we are able to predict involvement of various immune cell types in rheumatoid arthritis and less well-understood diseases such as

alopecia areata and Alzheimer's disease (Fig. 4i–k; see Supplementary Note 2 for further discussion).

Leukemogenesis and cancer evolution in AML

To characterize the evolution of AML[31] in the context of normal hematopoiesis, we identified 3 distinct stages of AML evolution: pre-leukemic HSCs (pHSCs), leukemia stem cells (LSCs), and leukemic blast cells (blasts) that can be enriched by FACS (Supplementary Fig. 11a,b). Current data indicate that HSCs serve as the reservoir for mutation acquisition during the early phases of leukemogenesis (Fig. 5a). Acquisition of founder mutations creates pHSCs that expand to create a pre-leukemic clone. Subsequent acquisition of progressor mutations generates LSCs that are capable of self-renewal and the production of AML blasts[32] (Fig. 5a).

Importantly, the population of HSCs isolated from leukemia patients by FACS represents a heterogeneous mixture of healthy unmutated HSCs and pHSCs. To quantify this heterogeneity, we define the "pre-leukemic burden" as the percentage of HSCs isolated from a leukemia patient that harbor at least the first mutation. We profiled the mutation frequency of known leukemogenic driver mutations in HSCs, T cells, and blast cells from 39 AML patients (Supplementary Table 5 and Supplementary Fig. 11c). Pre-leukemic burden is highly variable in this cohort with some patients exhibiting a complete repopulation of the HSC compartment with pre-leukemic cells and others exhibiting undetectable levels of pre-leukemic mutations (Fig. 5b, Supplementary Fig. 11d).

AML represents a cooption of normal myelopoiesis

The AML leukemogenic process provides a novel system to study the genesis and evolution of cancer. The Fast-ATAC protocol produced robust accessibility profiles from cryopreserved primary patient AML cells (Fig. 5c). We find that the level of variance in DNA accessibility between all samples of the same cell type increases through progressive stages of leukemia evolution (Fig. 5d, see methods). All AML cell types exhibit more inter-donor sample-to-sample variance than the corresponding normal hematopoietic cells (Fig. 5e). This may be a manifestation of the point along the normal hematopoietic hierarchy at which the particular AML cell types exist. Indeed, key developmentally-associated genes such as *GATA2* and *CEBPB* show variation amongst the AML cell types consistent with different developmental stages (Fig. 5f) and we find that the first four principal components derived from normal hematopoietic differentiation account for much of the variation observed in our leukemia samples (Fig. 5g, see methods). Assigning a score to the myeloid differentiation component of our data, we find that the various stages of AML spread across the trajectory from HSC to monocyte, indicating that the process of leukemogenesis largely mirrors the process of normal myelopoiesis (Fig. 5h and Supplementary Fig. 11e,f). Consistent with their functional ability to produce both lymphoid and myeloid cells in xenotransplantation assays[11–13], pHSCs are most closely related to HSCs and MPPs (Fig. 5h). As shown previously[33], LSCs show strong similarity to GMP and LMPP cells and leukemic blast cells show a wider distribution with less differentiated blasts clustering with GMP cells and more differentiated blasts clustering with monocyte cells[34,35] (Fig. 5h).

AML cell types exhibit regulatory heterogeneity

The observed developmental positions across myelopoiesis suggest that each patient-specific AML might harbor a unique collection of multiple distinct normal regulatory programs. Using enhancer cytometry, we quantified the contribution of each normal cell type to each leukemic sample assayed (Fig. 6a, Supplementary Fig. 12a, and Supplementary Table 6). We find that each patient, at each stage of leukemogenesis, harbors regulatory contributions from multiple distinct normal cell types that are often developmentally distinct from each other. This result raises the intriguing possibility that individual AML cells may either i) exist in mixed cell states that are not normally maintained during normal hematopoiesis, or ii) show cellular heterogeneity, wherein a mixture of cell states exist within the leukemic clone. Importantly, we find that the majority of the patient donors have AML blasts that are clonally derived and harbor all the leukemic mutations at comparable allele frequencies (Supplementary Table 5), suggesting that the epigenomic diversity observed through enhancer cytometry is not related to genetic heterogeneity of the AML cells.

To discriminate between these two possibilities, we performed single-cell ATAC-seq (scATAC-seq) on purified LSCs and blast cells from two AML patients and compared these samples to myeloid cells from healthy donors. We then performed enhancer cytometry using principal component analysis (PCA) trained on our ensemble ATAC-seq data (Fig. 6b; see methods). This analytical framework was validated by projection of down-sampled bulk ATAC-seq data (Supplementary Fig. 12b,c) and enabled accurate projection of single cell accessibility profiles onto hematopoietic principal components (Fig. 6c,d and Supplementary Fig. 12d,e). The relationship between developmental progression and single cell chromatin accessibility can be further visualized as a one-dimensional histogram (Fig. 6e,f and Supplementary Fig. 12e; see methods).

For normal physiologic comparison, we performed scATAC-seq on normal monocytes (N=88) and LMPPs (N=94) isolated from healthy donors. Single LMPP and monocyte cells show myelopoietic developmental projection scores centered at the predicted ensemble scores (Fig. 6e). In contrast, AML cells are either uniformly centered at developmentally intermediate states (e.g. SU070 LSC with unimodal peaks located between normal LMPPs and monocytes in Fig. 6f), or alternatively show broad bimodal distributions representing regulomes from intermediate and developmentally normal cell states (e.g. SU353 LSCs and blasts, Fig. 6f). In addition, widely used cell lines, such as the AML line HL60, also show a unimodal and mixed normal cell regulome, observed by ensemble and scATAC-seq (Supplementary Fig. 12f–j). These results show that the regulatory heterogeneity observed in the ensemble profiles of AML samples can arise from both single-cell intra- and inter-cellular heterogeneity (see Supplementary Note 3 for an extended discussion).

Synthetic normal analogs uncover AML-specific biology

The ability to accurately quantify the contribution of each normal cell regulome to the epigenetic profile of a leukemic cell type enables a more robust identification of AML-specific regulatory elements. In particular, analyses of leukemic cell types in the past have relied on comparing the malignant cells to a carefully chosen normal cell type (for example, GMP). Here, due to the regulatory heterogeneity in AML, we reasoned that an effective

normal cell comparison would be possible with the generation of “synthetic normals” which represent admixtures of various normal cells defined by enhancer cytometry (see methods). While comparison of AML cell types to their closest normal cell analogs yields a high correlation (R=0.80, Fig. 6g), comparison of AML cell types to their synthetic normal analogs yields a higher correlation (R=0.84, Fig. 6h and Supplementary Fig. 13a) and, more importantly, leads to a reduction in the number of AML-specific peaks identified (N=1,791 to N=899; Fig. 6i and Supplementary Fig. 13b,c). Also, comparing samples to the synthetic normal from each individual AML cell type reduces global measures of epigenetic variance (Supplementary Fig 13d compared to Fig. 5d).

To identify clusters of coordinately regulated elements, fold change values between each AML and its synthetic normal were clustered using k-means clustering to identify 7 distinct regulatory modules (Fig. 7a and Supplementary Fig. 14a; see methods). The usage of these modules was tracked through leukemogenesis to identify patterns related to specific AML cell types (Fig. 7b). Each module shows enrichment for peaks associated with different key transcription factors (Fig. 7c). For example, modules 6 and 7 show strong enrichment for JUN and FOS activity. Similar observations of increased JUN/FOS accessibility have been made from DNase-seq data in FLT3-ITD positive AML[20], suggesting that this result may be related to the high prevalence of FLT3 mutations in our patient cohort. This increase in accessibility of JUN/FOS motifs is reflected by an increase in expression of these factors by RNA-seq (Supplementary Fig. 14b) and is maintained through the stages of leukemogenesis, identifying inhibition of these pathways as a potential therapeutic strategy in AML (Supplementary Fig. 14c–e). This observation is consistent with previous publications that identify over-expression of *c-JUN* in AML[36] and find JNK inhibition as a putative therapeutic target[37,38] and indicates that similar strategies may prove efficacious in targeting pHSCs.

Mechanism and consequences of pHSC clonal advantage

Using ATAC-seq and enhancer cytometry we show that pHSCs share many regulatory programs with HSCs and MPPs (Fig. 6a). Nevertheless, comparison to synthetic normal analogs identifies distinct regulatory modules (modules 1 and 2) that show decreased accessibility in pHSCs, representing the earliest known event of AML evolution (Fig. 7b). These repressed regulatory modules are enriched for motifs associated with HSPCs (i.e. HOX, RUNX, and GATA) and provide direct evidence to support a model where pHSCs maintain a unique epigenetic and functional state.

In order to better understand the consequences of a loss in accessibility at motifs associated with HSPCs, we probed pHSCs for phenotypic changes related to self-renewal and differentiation. When pHSCs are induced to differentiate down the myeloid and erythroid lineages (Supplementary Fig. 14f), pHSCs showed a strong resistance towards differentiation, instead favoring maintenance of the stem cell immunophenotype as indicated by retention of CD34 expression (Fig. 7d,e). We hypothesized that the observed decreased accessibility at HOX transcription factor motifs might mediate the observed retention of stem cell immunophenotype. Indeed, depletion of one such HOX factor, HOXA9, by short hairpin RNA (shRNA) knockdown (Supplementary Fig. 14g and Supplementary Table 7) in

umbilical cord blood CD34+ HSPCs led to a retention of stem cell immunophenotype in the context of both myeloid (Fig. 7f) and erythroid (Fig. 7g) differentiation. Moreover, a concomitant decrease in differentiated granulocytes and erythroid cells was also observed (Supplementary Fig. 14h–j), consistent with results from mouse models of HOXA9 deficiency[39,40]. Together, these results suggest that decreased HOX accessibility in pHSCs may promote retention of stem cell characteristics and prevent differentiation of these cells. Additional HOX factors may play a role in defective pHSC differentiation, as the role of HOXA9 in hematopoiesis and leukemogenesis is complex[39–41].

pHSC resistance to differentiation potentially explains the observation that pHSCs outcompete their normal HSC counterparts *in vivo* (Supplementary Fig. 14k and Fig. 5b). pHSCs would gain an evolutionary advantage while promoting an HSC-like state, and thus increase the likelihood of acquiring additional leukemogenic mutations. One implication of this model is that pre-leukemic burden may have adverse effects on patient survival, despite the fact that pHSCs do not confer disease in xenograft transplant assays[11–13]. Characterization of our patient cohort shows that pre-leukemic burden inversely correlates with overall and relapse-free survival (hazard ratio = 3.30 for overall survival and 2.99 for relapse free survival, $p < 0.05$; Fig. 7h,i). These results further implicate pHSCs in AML pathology and suggest a mechanism whereby AML arises from a pre-leukemic clone that is capable of outcompeting its normal HSC counterparts (Supplementary Fig. 14k), which predisposes patients to more aggressive or refractory leukemia.

DISCUSSION

Here we report a rich resource charting the epigenomic and transcriptomic landscape of 16 unique blood cell types. This resource relies on the accurate and precise determination of the regulome landscapes in primary human blood cells, made possible by Fast-ATAC. Unsupervised clustering of accessible chromatin regions, specifically distal elements, groups individual cell types with high cluster purity (91% for ATAC-seq compared to 78% for RNA-seq), demonstrating that these distal regulatory elements more precisely define cell identity and developmental trajectory. Enhancer cytometry harnesses this specificity and enumerates the frequencies of pure cell types in complex cell mixtures. This technique may be applicable to address cell heterogeneity in other contexts of stem cell biology or cell therapy.

Additionally, this atlas of human hematopoiesis enriches the interpretation of GWAS results in several ways. We identify strong associations of disease-linked polymorphisms with the open chromatin landscapes of specific hematopoietic cell types, uncovering the developmental contexts in which the disease-relevant elements first become active. In the case of mean corpuscular volume, the strongest association occurs in erythroblast cells, but a significant association can be seen as early as the common myeloid progenitor stage (CMP). These results are consistent with the concept that many enhancers are developmentally primed prior to their activation following cell differentiation[3]. Our resource further provides a platform to identify specific trans-acting regulators that drive blood cell identity and function. Integration of ATAC-seq and RNA-seq data improves motif-transcription factor pairing and enables the accurate determination of causative regulators of chromatin

accessibility throughout hematopoietic differentiation. We anticipate this combined data set, which represents a dynamic developmental process, will be a rich resource for continued efforts to build computational tools that model both *cis*[42] and *trans*[43] determinants of chromatin accessibility and gene expression.

Application of this resource to the study of three distinct time points in AML evolution sheds light on the biology and step-wise progression of leukemia evolution. A longstanding debate in cancer biology is how cancer cells violate cell lineage rules[44,45], for example by maintaining self-renewal in an otherwise differentiated cell state. By using our comprehensive map of hematopoiesis, patient-matched AML cell subsets, and scATAC-seq of hundreds of individual leukemic and normal cells, we show evidence of regulatory heterogeneity in the epigenome—a single cell with several normally distinct regulatory programs (see supplementary discussion). We find that such mixed regulatory programs may be the result of both intra- and intercellular regulatory heterogeneity.

This regulatory heterogeneity demonstrates that there may be no appropriate “normal” for tumor–normal comparisons in epigenomic and transcriptomic studies. Instead, we use enhancer cytometry to construct “synthetic normals”—proportionally matching the predicted fractional contribution of cell type-specific regulomes from normal hematopoiesis —in order to pinpoint cancer-specific aberrations. This approach led us to identify the loss of HOX-mediated accessibility as the most consistent defect in pHSCs. We found that loss of a HOX factor can, in fact, cause defects in differentiation similar to those observed in pHSCs and potentially confer an evolutionary advantage. Importantly, higher pre-leukemic burden is predictive of poor overall and relapse-free survival in AML, indicating an important role for pHSCs in disease pathogenesis.

The methodologies developed here for the study of AML have important implications for the study of other blood and solid tumor malignancies. We anticipate that regulatory heterogeneity is a widespread phenomenon in many types of cancer, and that our integrative approach using enhancer cytometry to construct synthetic normal analogs should be broadly applicable to many disease pathologies. Future studies harnessing the power of enhancer cytometry to understand other cancer-specific regulatory networks will provide key insights into the aberrations that drive the formation and persistence of malignant disease. Thus, we believe that this work provides a methodological framework for the paradigm of mapping regulomes of normal tissues to better understand the ontogeny of human disease.

ONLINE METHODS

Availability of sequencing data

All sequencing data is available through the Gene Expression Omnibus (GEO) via accession GSE74912. Additionally, the data from normal hematopoietic cells has been made available as a UCSC Genome Browser Track Hub (see URLs) and as a Washington University EpiGenome Browser session (ID XVqu0IKMi1).

Human samples

Normal donor human bone marrow and peripheral blood cells were obtained fresh from AllCells (Alameda, CA) or the Stanford Blood Center (Palo Alto, CA). All normal blood cell populations were sorted fresh. Human AML samples were obtained from patients at the Stanford Medical Center with informed consent, according to Institutional Review Board (IRB)-approved protocols (Stanford IRB no. 18329 and 6453). Mononuclear cells from each sample were isolated by Ficoll separation, resuspended in 90% FBS + 10% DMSO, and cryopreserved in liquid nitrogen. All analyses conducted here on AML cells utilized freshly thawed cells. Criteria for inclusion of AML samples were pre-established. Samples were selected based solely on the availability of an adequate number of cells. For normal donors, no exclusion criteria were used.

Definition of cell types isolated

Here we isolate HSCs, LSCs, and blast cells from AML patients. These cells are defined by immunophenotype (Supplementary Table 1) as demonstrated previously[46]. The patients examined by ATAC-seq and RNA-seq in this study were selected in such a way that >80% of the HSCs are pre-leukemic.

Additionally, we isolate multiple different normal cell types from healthy donors (Supplementary Table 1). Mature granulocytes were excluded from our analyses due to high endogenous RNases and proteases. Mature megakaryocytes proved difficult to isolate in adequate cell numbers and were similarly excluded.

Cell lines

Cell line data was downloaded from GEO accession number GSE65360.

Flow cytometry analysis and cell sorting

All antibodies used for flow cytometry are detailed in Supplementary Table 1).

To prepare cells for FACS, all cells were recovered for 20 minutes at 37°C in the presence of 200 U/ml DNase (Worthington Biochemical, Lakewood, NJ) in IMDM with 10% fetal bovine serum. After recovery, viable mononuclear cells were separated by a Ficoll density gradient (GE Healthcare). When necessary, CD34-based enrichment was performed using paramagnetic MACS beads (Miltenyi Biotech Inc, San Diego, CA) per the manufacturer's protocol.

FACS sorting was performed on a Becton Dickinson FACSAria II. All cells were resuspended in and sorted into cold FACS Buffer (PBS + 2% FBS + 2 mM EDTA) containing propidium iodide at 1 ug/ml or 4′,6-diamidino-2-phenylindole (DAPI) at 1 ug/ml. All cell sorting steps were validated using post-sort analyses to verify purity of sorted cell populations (Supplementary Table 1).

Transcriptome sequencing

RNA was isolated from 1,000–100,000 FACS-purified cells using the Qiagen RNeasy Plus Micro Kit. RNA quality was verified on an Agilent Bioanalyzer Pico Eukaryote chip. 5 ul of

total RNA (300 pg – 80 ng) was used as input into the NuGen Ovation V2 cDNA synthesis kit. SPIA-amplified cDNA was sheared using a Covaris S2 sonicator as follows: 10% duty cycle, 5 intensity, 100 cycles/burst, 5 minutes, 120 ul volume. Sheared cDNA was purified and size selected using Ampure XP beads at a 0.9:1 beads:sample ratio. After cleanup, Illumina TruSeq adapters were ligated onto the cDNA using the NEB Next Ultra library prep kit per manufacturer instructions. Library quality and concentration were determined using an Agilent Bioanalyzer HS DNA chip and a Qubit fluorometer. Libraries were sequenced to an average depth of 12 million read pairs per sample.

Transcriptome data analysis

RNA sequencing data was aligned to the human reference genome (GRCh37/hg19) using STAR using standard input parameters. Aligned reads were filtered for those reads that map uniquely to non-mitochondrial regions. Duplicate reads were removed using PICARD MarkDuplicates. Transcript counts were produced using HTseq against the UCSC refGene transcriptome. Transcript counts were processed using DESeq2, normalizing for both library size and transcript GC content using Conditional Quantile Normalization[47]. Differential expression was determined without the use of a Cooks cutoff. All downstream analyses on RNA-seq data were performed on variance stabilizing transformed data obtained from DESeq2.

Fast-ATAC sequencing

This protocol has been optimized for blood cells. We note that digitonin is a gentle detergent and this protocol may not be ideal for cell lines and other cell types that are more resistant to lysis. 5,000 sorted cells in FACS Buffer were pelleted by centrifugation at 500 RCF for 5 minutes at 4C in a pre-cooled fixed-angle centrifuge. All supernatant was removed using two pipetting steps being careful to not disturb the not visible cell pellet. 50 ul transposase mixture (25 ul of 2x TD buffer, 2.5 ul of TDE1, 0.5 ul of 1% digitonin, 22 ul of nuclease-free water) (Cat# FC-121-1030, Illumina; Cat# G9441, Promega) was added to the cells and the pellet was disrupted by pipetting. Transposition reactions were incubated at 37°C for 30 minutes in an Eppendorf ThermoMixer with agitation at 300 RPM. Transposed DNA was purified using a QIAgen MinElute Reaction Cleanup kit (Cat# 28204) and purified DNA was eluted in 10 ul elution buffer (10 mM Tris-HCl, pH 8). Transposed fragments were amplified and purified as described previously[48] with modified primers[23]. Libraries were quantified using qPCR prior to sequencing. All Fast-ATAC libraries were sequenced using paired-end, dual-index sequencing on a NextSeq with 76×8×8×76 cycle reads.

ATAC-seq data analysis

ATAC-seq data were processed as previously described[23] with notable exceptions. In brief, reads were trimmed using a custom script and aligned using Bowtie2. To call peaks, data were aggregated by each unique cell type, peak summits were called using MACS2, and filtered using a custom blacklist, as previously described[23].

To generate a non-redundant list of hematopoiesis and cancer peaks we first extended summits to 500 bp windows (+/− 250 bps). We then ranked 500 bp peaks by their summit significance value (defined by MACS2) and chose a list of non-overlapping, maximally

significant, peaks. The complete data set comprised a total of 590,650 peaks. To annotate peaks with promoter/distal labels, and nearest gene, we used the Homer package, with the command "annotatePeaks.pl". As described previously[23], we counted fragments for each sample across all 590,650 peaks to provide a count matrix. To obtain normalized fragment counts, which were used for all downstream processing, we first performed quantile normalization followed by GC normalization (CQN R package[47]). Data tracks, used solely for visualization, were normalized to the number of fragments falling within all peaks for each sample. Coverage tracks were visualized using the Gviz R-package. Fragment yield (Supplementary Fig. 1e), was computed by multiplying the library diversity calculated using PICARD tools with the number of reads falling within peaks, values were then divided by the number of cells used in each assay.

For information on TF-based analyses, see Supplementary Note 1.

Unsupervised hierarchical clustering

Unless otherwise stated, all hierarchical clustering was unsupervised using Pearson correlation as the distance metric and performed on all relevant features (for ex. all genes for RNA-seq or all peaks for ATAC-seq). All clustering analyses were performed on normalized data as described in the relevant methods sections.

Cluster Purity

Cluster purity is calculated as described previously[19]. Briefly, 13 clusters were defined as the branches of the dendrogram that represent all individual replicates without overlap. Each cluster is assigned to the cell type which is most frequent in the cluster. In this way, there is one cluster (branch) that is assigned to represent each cell type. For each cluster, the accuracy of this assignment is measured by counting the number of correctly assigned experiments. For example, if the "HSC cluster" contained 3 HSC experiments and 2 MPP experiments, this cluster would be given a value of 3. The sum of correctly assigned experiments is divided by the total number of experiments to give the cluster purity.

GWAS Analysis

Using a list of blood-enriched GWAS, we applied the "deviation" pipeline (as described in the previous section for TF motifs), using an identical approach wherein each GWAS disease is analogous to a TF motif and each GWAS peak association is analogous to an individual TF motif occurrence in a peak. For more information, see Supplementary Note 1.

CIBERSORT application, benchmarking, and signature matrix generation

CIBERSORT v1.0.1 was used as recommended by the authors. Test set data and training set data represented unique non-overlapping samples. Benchmarking was performed using randomly permuted synthetic data. For each test, a unique signature matrix was made from N–1 replicates of each cell type ("leave-one-out"). This signature matrix was used to deconvolve 10 randomly permuted cellular mixtures derived from the replicate that was excluded from the training set and signature matrix. One hundred unique permutations were performed, 10 permutations each on 10 different training sets.

The curated CIBERSORT signature matrix (Supplementary Table 2) was generated using the default CIBERSORT parameters. To define a list of distal elements for input into CIBERSORT, we filtered peaks by removing peaks mapping to sex chromosomes, promoter/TSS regions (+/− 1 kb), and regions found to be highly accessible in AML samples when compared to the closest normal cell-type. Artefactual peaks were also removed using a custom blacklist as described above. These regions were removed to prevent bias based on donor gender, enhance cell type-specific patterns, and avoid over-fitting of AML samples to normal cell types respectively.

Generation of synthetic normal analogs

Synthetic normal analogs were generated based on the fractional contributions predicted by CIBERSORT (Supplementary Table 6). For each AML sample, a synthetic normal analog was generated by multiplying the fractional contribution of each normal cell type by the normalized fragment number for that cell type. This is done on a peak-by-peak basis and the values are summed for each peak to give the synthetic normal value. For example, assuming a given sample has a fractional contribution of 0.3 HSC, 0.5 MPP, 0.2 CMP, and 0 for all other cell types: a synthetic normal analog for peak #1 would be constructed by taking the sum of the average HSC normalized fragments multiplied by 0.3, MPP multiplied by 0.5, CMP multiplied by 0.2, and all other cell types by zero. Synthetic normal analogs were then quantile normalized with the leukemic sample of interest.

Cancer modules

Synthetic normal analogs for each cancer sample were generated as described above. To calculate differences between tumor-synthetic normal pairs we computed log2(fold change) values from the AML sample of interest to the corresponding synthetic normal. Importantly, samples SU209-pHSC and SU583-pHSC were removed from this analysis. These samples appeared to be outliers in that they were more developmentally mature and exhibited an unexpectedly large number of differential peaks (Supplementary Fig. 13b). To determine unique cancer-specific regulatory modules, we first filtered for significantly altered peaks using a cutoff of log2(fold change) greater than 4 or less than −4 resulting in 6,752 peaks. To determine AML-specific regulatory modules, we used k-means to cluster the significantly altered peaks, described above. A K=7 was determined by analyzing the mean centroid distances of each cluster (Euclidean) for an increasing K from 1 to 20 (sSupplementary Figure 14a) where a K=7 approximated much of the peak dynamics observed. To determine motif enrichments within each module, we calculate the fraction of motif instances in a given module peak set and divide by all motif instances in all observed peaks.

AML sample genotyping

All AML patient samples described here were genotyped either by whole exome sequencing using the SeqCap EZ Exome SR kit v3.0 (Roche/Nimblegen) or by customized hybrid capture sequencing of the 130 genes most frequently mutated in AML[49] (see methods) using the SeqCap EZ Choice kit (Roche/Nimblegen). Sequencing was performed on an Illumina HiSeq 2000, HiSeq 2500, or NextSeq 500. Sequence data were aligned to the human reference genome hg19 using BWA (v0.5.9) for global alignment and GATK (v2.8-1) for local realignment. Aligned reads were processed for downstream mutation calling using

SAMTools (v0.1.12a). SNPs were called using GATK and Varscan (v2.3.7). All data derived from customized hybrid capture did not have a matched normal genome and was compared instead to the hg19 human reference genome. Putative SNPs were filtered for: 1) minimum sequence depth of 50 reads, 2) less than 90% variant strand bias, 3) non-synonymous, 4) if the SNP is observed in dbSNP, the MAF must be less than 1%, 5) minimum variant frequency of 5%. Insertions and deletions (indels) were called using GATK[50] and Varscan[51]. Putative indels were filtered for: 1) minimum sequence depth of 25, 2) minimum variant frequency of 5%, 3) less than 90% variant strand bias, 4) not observed in dbSNP. Large-scale genomic events such as translocations were called using FACTERA[52] (v1.3) with no additional filtering. FLT3 internal tandem duplications were called using Pindel[53] (v0.2.4) with no additional filtering. Manual observation was used to clarify borderline mutation calls. Additional weight was given to mutations called by more than one algorithm. All mutations were validated by targeted amplicon sequencing.

Targeted amplicon sequencing of leukemia-associated mutations

Targeted Amplicon Sequencing was performed as described previously[12].

Epigenetic variance calculation

Epigenetic variance was calculated as the sum of the squares of the distance from the mean divided by the number of samples. This is equivalent to the VAR.P function in Microsoft Excel. This variance was calculated for each individual peak. To obtain the genome wide variance the rolling mean of 10,000 sequential peaks was calculated across the linear genome in chromosomal order. For calculations of epigenetic variance some samples with high background were omitted.

Analysis of DNase data

DNase CD34 data, made available by the Epigenomics Roadmap Consortium, was downloaded from SRA accession numbers SRR066150, SRR066151, SRR066152, SRR066351, SRR097542, SRR327476, SRR327477. Single-end DNase data was aligned, filtered and normalized using the methods described in the ATAC-seq data processing section.

Correlating TF motif deviation scores to expressed genes

Genes were first filtered for putative transcription factors (N=1,820)[54]. log2(fold change) and standard error on the mean (SEM) were computed using DESeq2 (as described above). To determine robust correlation coefficients (Pearson) and p-values for genes and TF deviation scores (as described above), we permuted (N=1,000) log2(fold change) values according to the measurement error as determined by SEM. Reported Pearson correlation coefficients represent the mean across the sampled data. Reported p-values represent a z-test statistic across the permutations.

To determine putative direct regulators of the given motif, we downloaded all available *in vitro* and inferred PWMs from CIS-BP[55]. We then calculated correlation coefficients (Pearson) of all CIS-BP PWMs (N=7,592) with the unique set of hematopoiesis PWMs (N=46). To account for offsets we take the maximum calculated correlation coefficient after

aligning two motifs in both orientations (reverse complement) and all possible offsets of length K. To filter the complete CIS-BP database (N=7,592) to a non-redundant gene list (N=806), we choose the motif with the maximum similarity (Pearson) to any hematopoiesis TF motif (Supplementary Fig. 9b and Supplementary Table 4). To find putative direct regulators of human hematopoiesis we filtered for TFs with a PWM correlation coefficient >0.8 (Supplementary Fig. 9e). Although we find many TFs can be correlated with their motif usage, we report the most correlated TF (Supplementary Fig. 9g,h) and the complete list in Supplementary Table 4.

Single-cell ATAC-seq analysis and enhancer cytometry

Single-cell ATAC-seq and enhancer cytometry analysis were performed as described in Supplementary Note 1.

Survival analysis

Overall survival was defined as the time from diagnosis to death from any cause. Relapse-free survival was defined as the time from complete morphologic remission to date of relapse of AML or death from any cause, whichever came first. Survival analysis was performed using the Kaplan-Meier estimate method. All patients were included for the analysis regardless of their treatment. P values comparing two Kaplan-Meier survival curves were calculated using the log-rank (Mantel-Cox) test. Hazard ratios were determined using the Mantel-Haenszel approach.

In vitro culture of primary AML cells for drug sensitivity

Primary AML blasts were cultured in Myelocult H5100 (Stemcell Technologies) with 20 ng/ml FLT3L, SCF, TPO, IL3, IL6 and 0.5 ug/ml Hydrocortisone. Blasts were cultured at 1 million cells/ml for a total of 6 days with no media changes. Drug sensitivity was measured by flow cytometric analysis of annexin negative, DAPI negative cells, live cells.

In vitro culture assays on HSPCs

FACS-purified HSPCs were plated into either myeloid differentiation media [Myelocult H5100 (Stemcell Technologies) with 20 ng/ml FLT3L, IL3, TPO, SCF, and GM-CSF, and 0.5 ug/ml Hydrocortisone] or erythroid differentiation media [StemSpan SFEM II (Stemcell Technologies) with the Erythroid Expansion Supplement (Stemcell Technologies)] and cultured for 6 days with media changes as necessitated by cellular proliferation. Stemness retention media is HPGM (Lonza) containing 20 ng/ml FLT3L, SCF, and TPO.

Knockdown of HOXA9

HOXA9 knockdown was achieved using the pRSI9 lentiviral backbone (Cellecta) that allows for constitutive expression of shRNA from a U6 promoter. The shRNA target sequences can be found in Supplementary Table 7.

IC50 determination in primary AML cells

Cell death in response to pharmacologic inhibition was measured by Annexin V staining using an Annexin V – AlexaFluor 647 conjugate (Life Technologies) as per the

manufacturer's instructions. Responses were measured in relation to a vehicle-treated control.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Claire Mazumdar and Anil Raj for assistance with RNA-seq, Aaron Newman for expert assistance with CIBERSORT, and our lab members for discussion. We thank the Stanford Hematology Division Tissue Bank and the patients for donating their samples. M.R.C. acknowledges NIH training grant R25CA180993 and NIH F31 Predoctoral fellowship F31CA180659. J.D.B. acknowledges the National Science Foundation Graduate Research Fellowships and NIH training grant T32HG000044 for support. M.P.S. acknowledges the NIH and the National Human Genome Research Institute (NHGRI) for funding through 5U54HG00455805. Supported by NIH (P50-HG007735 to H.Y.C., W.J.G., M.P.S.), UH2-AR067676 (H.Y.C), Stanford Cancer Center (H.Y.C.), HHMI (H.Y.C., J.K.P.), Stinehart-Reed Foundation (R.M.), Ludwig Institute (R.M.), NIH (R01CA18805 to R.M.). R.M. is a New York Stem Cell Foundation Robertson Investigator.

References

1. Quesenberry, PJ.; Colvin, GA. Williams Hematology. McGraw-Hill; 2005. Hematopoietic Stem Cells, Progenitor Cells, and Cytokines.
2. Ji H, et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature. 2010; 467:338–342. [PubMed: 20720541]
3. Lara-Astiaso D, et al. Chromatin state dynamics during blood formation. Science. 2014; 55:1–10.
4. Chen L, et al. Transcriptional diversity during lineage commitment of human blood progenitors. Science. 2014; 345:1251033–1251033. [PubMed: 25258084]
5. Novershtern N, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011; 144:296–309. [PubMed: 21241896]
6. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013; 10:1213–1218. [PubMed: 24097267]
7. Jin W, et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature. 2015; 528:142–6. [PubMed: 26605532]
8. Shih AH, Abdel-Wahab O, Patel JP, Levine RL. The role of mutations in epigenetic regulators in myeloid malignancies. Nat Rev Cancer. 2015; 263:22–35.
9. Lindberg J, et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. N Engl J Med. 2014; 371:2477–2487. [PubMed: 25426838]
10. Jaiswal S, et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. N Engl J Med. 2014; 371:2488–2498. [PubMed: 25426837]
11. Jan M, et al. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. Sci Transl Med. 2012; 4:1–10.
12. Corces-Zimmerman MR, Hong WJ, Weissman IL, Medeiros BC, Majeti R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. Proc Natl Acad Sci U S A. 2014; 111:2548–53. [PubMed: 24550281]
13. Shlush LI, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature. 2014; 506:328–333. [PubMed: 24522528]
14. Majeti R, Park CY, Weissman IL. Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. Cell Stem Cell. 2007; 1:635–45. [PubMed: 18371405]
15. Manz MG, Miyamoto T, Akashi K, Weissman IL. Prospective isolation of human clonogenic common myeloid progenitors. Proc Natl Acad Sci U S A. 2002; 99:11872–11877. [PubMed: 12193648]

16. Kohn, La, et al. Lymphoid priming in human bone marrow begins before expression of CD10 with upregulation of L-selectin. Nat Immunol. 2012; 13:963–971. [PubMed: 22941246]
17. Seita J, Weissman IL. Hematopoietic stem cell: self-renewal versus differentiation. Wiley Interdiscip Rev Syst Biol Med. 2010; 2:640–653. [PubMed: 20890962]
18. Roadmap Epigenetics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
19. Manning, CD.; Raghavan, P.; Schutze, H. Introduction to Information Retrieval. Cambridge University Press; 2008.
20. Cauchy P, et al. Chronic FLT3-ITD Signaling in Acute Myeloid Leukemia Is Connected to a Specific Chromatin Signature. Cell Rep. 2015; 12:821–836. [PubMed: 26212328]
21. Heinz S, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. 2010; 38:576–589. [PubMed: 20513432]
22. Newman AM, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015; 12:1–10. [PubMed: 25699311]
23. Buenrostro JD, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015; 523:486–490. [PubMed: 26083756]
24. Weiss MJ, Orkin SH. GATA transcription factors: key regulators of hematopoiesis. Exp Hematol. 1995; 23:99–107. [PubMed: 7828675]
25. Burns CE, Traver D, Mayhall E, Shepard JL, Zon LI. Hematopoietic stem cell fate is established by the Notch-Runx pathway. Genes Dev. 2005; 19:2331–42. [PubMed: 16166372]
26. Nerlov C, Graf T. PU. 1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. Genes Dev. 1998; 12:2403–2412. [PubMed: 9694804]
27. Neph S, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012; 489:83–90. [PubMed: 22955618]
28. Gjoneska E, et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature. 2015; 518:365–369. [PubMed: 25693568]
29. Farh KK, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015; 518:337–343. [PubMed: 25363779]
30. Maurano MT, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012; 337:1190–1195. [PubMed: 22955828]
31. Dohner H, Weisdorf DJ, Bloomfield CD. Acute Myeloid Leukemia. N Engl J Med. 2015; 373:1136–52. [PubMed: 26376137]
32. Bonnet D, Dick JE. Huamn acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nat Med. 1997; 3:730–737. [PubMed: 9212098]
33. Goardon N, et al. Coexistence of LMPP-like and GMP-like Leukemia Stem Cells in Acute Myeloid Leukemia. Cancer Cell. 2011; 19:138–152. [PubMed: 21251617]
34. Bennet JM, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. Br J Haematol. 1976; 33:451–8. [PubMed: 188440]
35. van't Veer MB. The diagnosis of acute leukemia with undifferentiated or minimally differentiated blasts. Ann Hematol. 1992; 64:161–5. [PubMed: 1581403]
36. Rangatia J, et al. Elevated c-Jun expression in acute myeloid leukemias inhibits C/EBPalpha DNA binding via leucine zipper domain interaction. Oncogene. 2003; 22:4760–4764. [PubMed: 12879022]
37. Volk A, et al. Co-inhibition of NF- B and JNK is synergistic in TNF-expressing human AML. J Exp Med. 2014; 211:1093–1108. [PubMed: 24842373]
38. Hartman AD, et al. Constitutive c-jun N-terminal kinase activity in acute myeloid leukemia derives from Flt3 and affects survival and proliferation. Exp Hematol. 2006; 34:1360–1376. [PubMed: 16982329]
39. Magnusson M, Brun ACM, Lawrence HJ, Karlsson S. Hoxa9/hoxb3/hoxb4 compound null mice display severe hematopoietic defects. Exp Hematol. 2007; 35:1421.e1–1421.e9. [PubMed: 17761289]

40. Lawrence HJ, et al. Mice bearing a targeted interruption of the homeobox gene HOXA9 have defects in myeloid, erythroid, and lymphoid hematopoiesis. Blood. 1997; 89:1922–1930. [PubMed: 9058712]
41. Thorsteinsdottir U, et al. Overexpression of the myeloid leukemia – associated Hoxa9 gene in bone marrow cells induces stem cell expansion. 2002; 99:121–129.
42. González AJ, Setty M, Leslie CS. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. Nat Genet. 2015; 47:1249–1259. [PubMed: 26390058]
43. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nat Methods. 2015; 12:265–272. [PubMed: 25240437]
44. Macedo A, et al. Characterization of aberrant phenotypes in acute myeloblastic leukemia. Ann Hematol. 1995; 70:189–194. [PubMed: 7748963]
45. Tiacci E, et al. PAX5 expression in acute leukemias: Higher B-lineage specificity than CD79a and selective association with t(8;21)-acute myelogenous leukemia. Cancer Res. 2004; 64:7399–7404. [PubMed: 15492262]
46. Jan M, et al. Prospective separation of normal and leukemic stem cells based on differential expression of TIM3, a human acute myeloid leukemia stem cell marker. Proc Natl Acad Sci U S A. 2011; 108:5009–14. [PubMed: 21383193]
47. Hansen KD, Irizarry Ra, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012; 13:204–216. [PubMed: 22285995]
48. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015; 109:21.29.1–21.29.9. [PubMed: 25559105]
49. TCGA Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013; 368:2059–74. [PubMed: 23634996]
50. McKenna A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]
51. Koboldt DC, et al. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009; 25:2283–2285. [PubMed: 19542151]
52. Newman AM, et al. FACTERA: a practical method for the discovery of genomic rearrangements at breakpoint resolution. Bioinformatics. 2014; 30:3390–3. [PubMed: 25143292]
53. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25:2865–2871. [PubMed: 19561018]
54. Vaquerizas JM, Kummerfeld SK, Teichmann Sa, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009; 10:252–263. [PubMed: 19274049]
55. Weirauch MT, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell. 2014; 158:1431–1443. [PubMed: 25215497]
56. Leslie R, O'Donnell CJ, Johnson aD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics. 2014; 30:i185–i194. [PubMed: 24931982]
57. Barrett JC, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009; 41:703–707. [PubMed: 19430480]
58. Lambert JC, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013; 45:1452–1458. [PubMed: 24162737]
59. De Vita S, et al. Efficacy of selective B cell blockade in the treatment of rheumatoid arthritis: evidence for a pathogenetic role of B cells. Arthritis Rheumatol. 2002; 46:2029–33.
60. Coenen MJH, Gregersen PK. Rheumatoid arthritis: a view of the current genetic landscape. Genes Immun. 2009; 10:101–111. [PubMed: 18987647]
61. Petukhova L, et al. Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. Nature. 2010; 466:113–117. [PubMed: 20596022]

62. Butovsky O, Kunis G, Koronyo-Hamaoui M, Schwartz M. Selective ablation of bone marrow-derived dendritic cells increases amyloid plaques in a mouse Alzheimer's disease model. Eur J Neurosci. 2007; 26:413–416. [PubMed: 17623022]

63. El Khoury J, et al. Ccr2 deficiency impairs microglial accumulation and accelerates progression of Alzheimer-like disease. Nat Med. 2007; 13:432–438. [PubMed: 17351623]

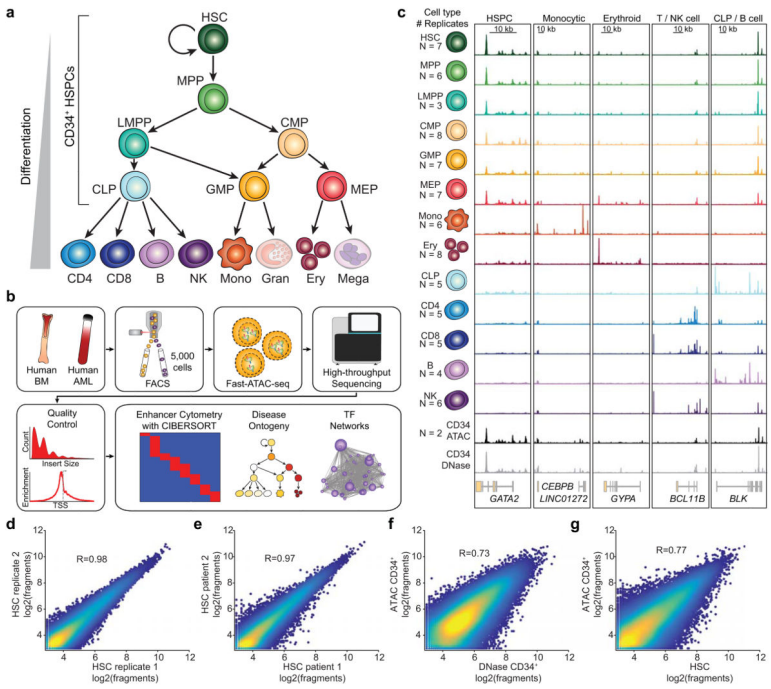

Figure 1. Interrogation of chromatin landscapes in primary blood cells

(a) Schematic of the human hematopoietic hierarchy shows the 13 primary cell types analyzed in this work. Granulocytes and megakaryocytes were excluded. The cell types comprising the CD34+ HSPCs are indicated. Colors used in this schematic are consistent throughout the manuscript.

(b) Diagram of analyses performed using paired ATAC-seq and RNA-seq data in both primary human blood cells and primary patient AML cells.

(c) Normalized ATAC-seq profiles at developmentally important genes. Profiles represent the union of all technical and biological replicates for each cell type. See Supplementary Table 1 for the exact number of technical and biological replicates for each cell type. Genomic coordinates of regions: GATA2 chr3:128,197,777–128,218,433; CEBPB chr20:48,800,260–48,904,715; GYPA chr4:145,020,689–145,070,000; BCL11B chr14:99,513,898–99,796,947; BLK chr8:11,343,117–11,429,285. All Y axis scales range from 0–10 in normalized arbitrary units. X axis scales indicated by scale bar.

(d–g) Scatter plot showing correlation of (d) technical replicates, (e) different human donors, (f) ATAC-seq and DNase-seq data derived from CD34+ HSPCs, and (g) ATAC-seq HSCs with bulk CD34+ HSPCs. The R values reported are calculated from correlations of all peaks. Plots show 50,000 random peaks, each with at least 5 reads.

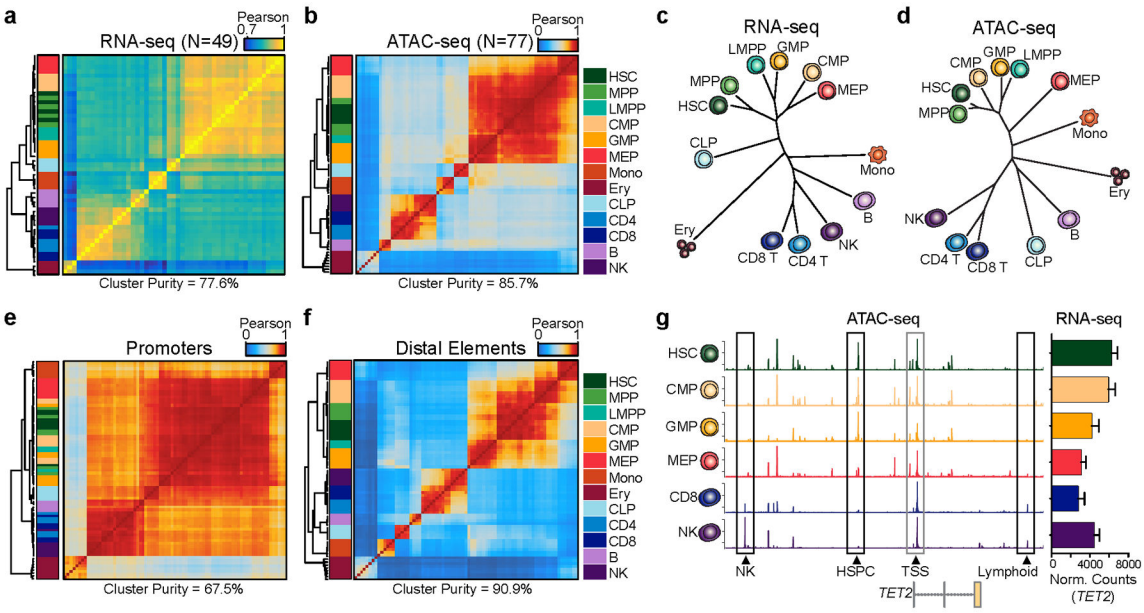

Figure 2. Distal regulatory elements enable accurate classification of the hematopoietic hierarchy

(a,b) Unsupervised hierarchical clustering of (a) RNA-seq (N=49) and (b) ATAC-seq (N=77) data from all replicates of 13 normal hematopoietic cell types. Values shown are Pearson correlation coefficients. Cluster purity quantifies the degree that cells of the same lineage (color coded in the key) are clustered together. For RNA-seq, clustering was performed using variance stabilizing transform-normalized expression data for all expressed annotated genes. For ATAC-seq, clustering was performed on all peaks using quantile normalized quantitative read coverage data.

(c,d) Phylogenetic dendrograms of (c) RNA-seq and (d) ATAC-seq data showing inter-cell type correlations derived from aggregate averages of all biological and technical replicates. Length of tree branches represents Euclidean distance. Data represents the union of all technical and biological replicates for each cell type.

(e,f) Hierarchical clustering of ATAC-seq profiles (N=77) mapping to (e) promoters and (f) distal regulatory elements. Values shown are Pearson correlation coefficients. Promoter-proximal peaks are defined as +/−1 kilobase from an annotated TSS. Distal element peaks are defined as those peaks greater than 1 kilobase from an annotated TSS.

(g) ATAC-seq peaks in the *TET2* locus show highly variable distal regulatory landscapes (left) and relatively constitutive expression of *TET2* (right). Data represents the union of all technical and biological replicates for each cell type: HSC=7; CMP=8; GMP=7; MEP=7; CD8=5; NK=6. Error bars represent 1 standard deviation. Genomic coordinates: chr4:106031731–106073198. Y axis scale ranges from 0–10 in normalized arbitrary units.

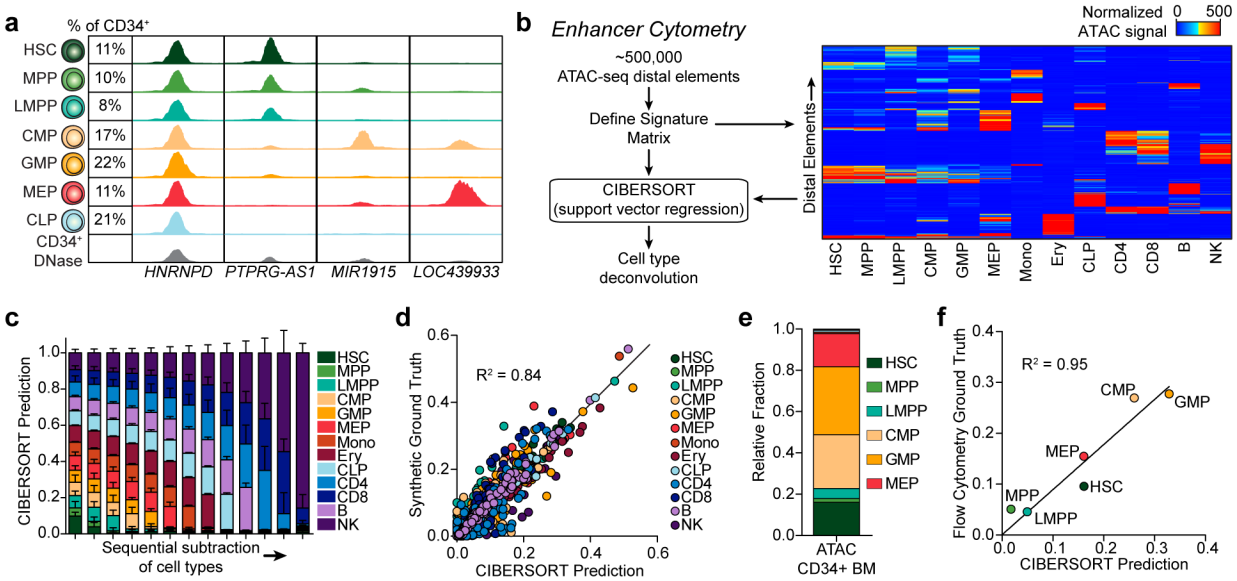

Figure 3. Enhancer cytometry allows for deconvolution of the hematopoietic hierarchy

(a) Normalized ATAC-seq profiles of HSPC subsets and ensemble CD34+ HSPC DNase-seq profiles illustrating heterogeneity amongst CD34+ HSPC subpopulations. Predicted cell fractions based on flow cytometry of 6 healthy bone marrow donors are shown on the left and the nearest annotated genes are shown on the bottom. Genomic coordinates: PTPRG-AS1 chr3:62194000–62196000; LOC439933 chr4:35761750–35763750; MIR1915 chr10:21639750–21641750; HNRNPD chr4:83205250–83207250. Y axis scale ranges from 0–10 in normalized arbitrary units.

(b) Schematic of enhancer cytometry from cell-type specific distal elements (right panel, N=735). The signature matrix heatmap has an upper threshold of 500 where all elements with signal greater than 500 appear red.

(c,d) Benchmarking of enhancer cytometry using randomly permuted synthetic mixtures to test robustness to (c) sequential subtraction and (d) randomized mixture content. Test data and training data are non-overlapping. (c) Synthesized ground truths are equal mixtures of the remaining cell types. In the left-most column, all cell types are present in equal parts in the ground truth data. Cell types are then sequentially subtracted from the synthesized ground truth starting with HSC until only NK cells remain. Error bars represent the standard deviation of 100 random permutations.

(e) Enhancer cytometry of ATAC-seq data derived from FACS-purified bone marrow CD34+ HSPCs.

(f) Correlation of predicted fractional contribution of each HSPC cell type by enhancer cytometry versus flow cytometric “ground truth” data of input CD34+ cells. X-axis represents the same data as shown in (e).

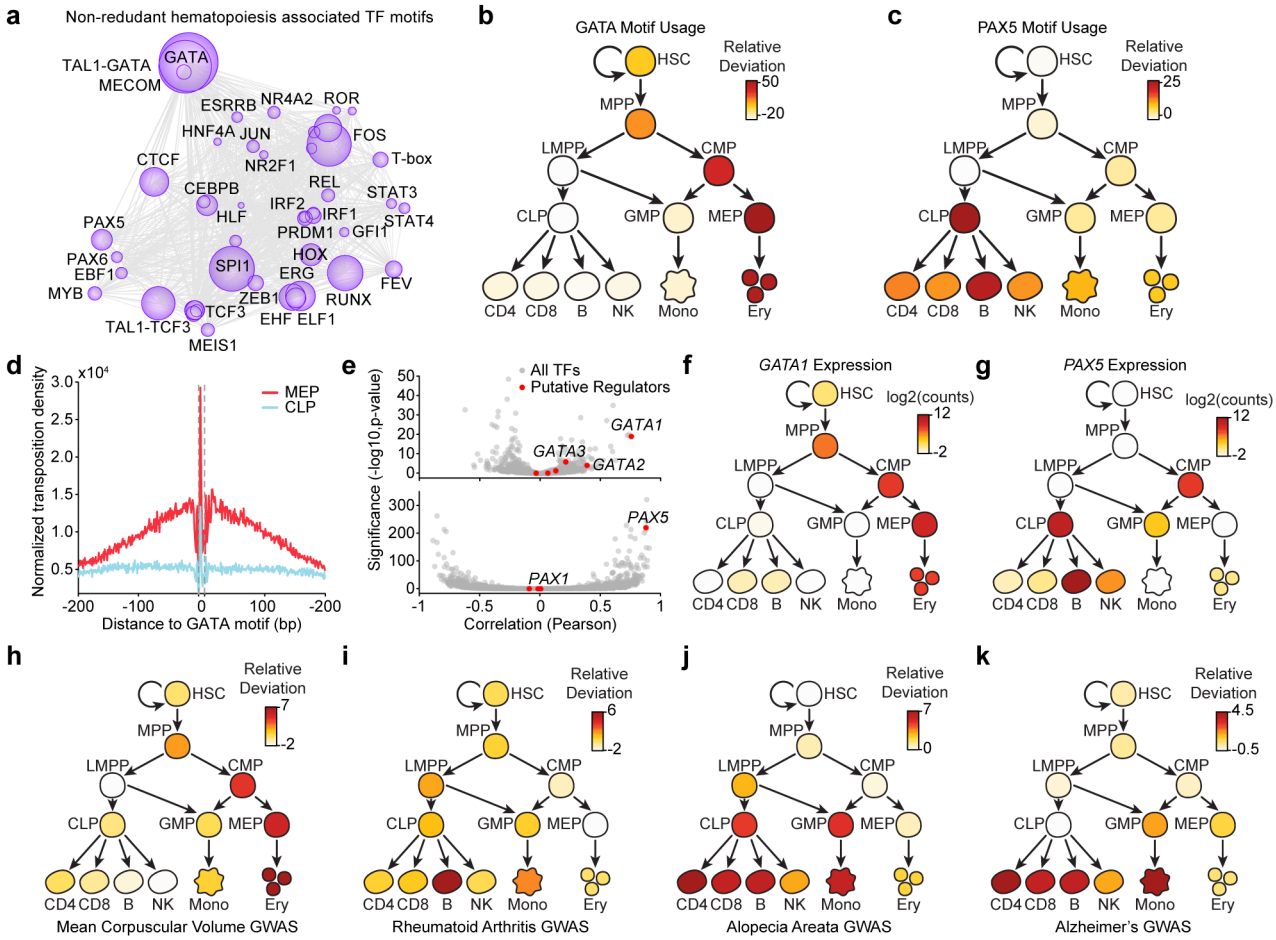

Figure 4. Integrative analysis of the hematopoietic regulome refines transcriptional circuitry driving cell specification and enriches the understanding of human disease

(a) Transcription factor dynamics showing major TFs driving hematopoietic regulomes. The size of the circle represents the effect of that motif in driving accessibility in human blood cells. The relative distance between circles represents the co-occurrence of motifs throughout hematopoietic differentiation (see methods).

(b,c) Usage of the (b) GATA and (c) PAX motif throughout hematopoietic differentiation. Values represent the relative deviation of the motif accessibility, a measure of motif usage, compared to that in HSCs.

(d) Footprint analysis of the GATA motif in MEP and CLP cells.

(e) Pearson correlation of motif accessibility with transcription factor expression plotted against the significance of this correlation for GATA (top) and PAX (bottom) motifs. Red dots represent DNA-binding factors found in the analysis in Supplementary Fig. 9b to bind the given motif. Gray dots represent all other DNA-binding factors.

(f,g) Expression of (f) *GATA1* and (g) *PAX5* phenocopies the usage of the (b) GATA and (c) PAX motifs throughout hematopoietic differentiation

(h–k) Relative deviation scores of chromatin accessibility within hematopoietic regulatory elements with GWAS SNPs for (h) mean corpuscular volume, (i) rheumatoid arthritis, (j) alopecia areata, and (k) Alzheimer's disease (see methods). Darker red color is

representative of enrichment of GWAS SNPs in the open chromatin regions of the given cell type.

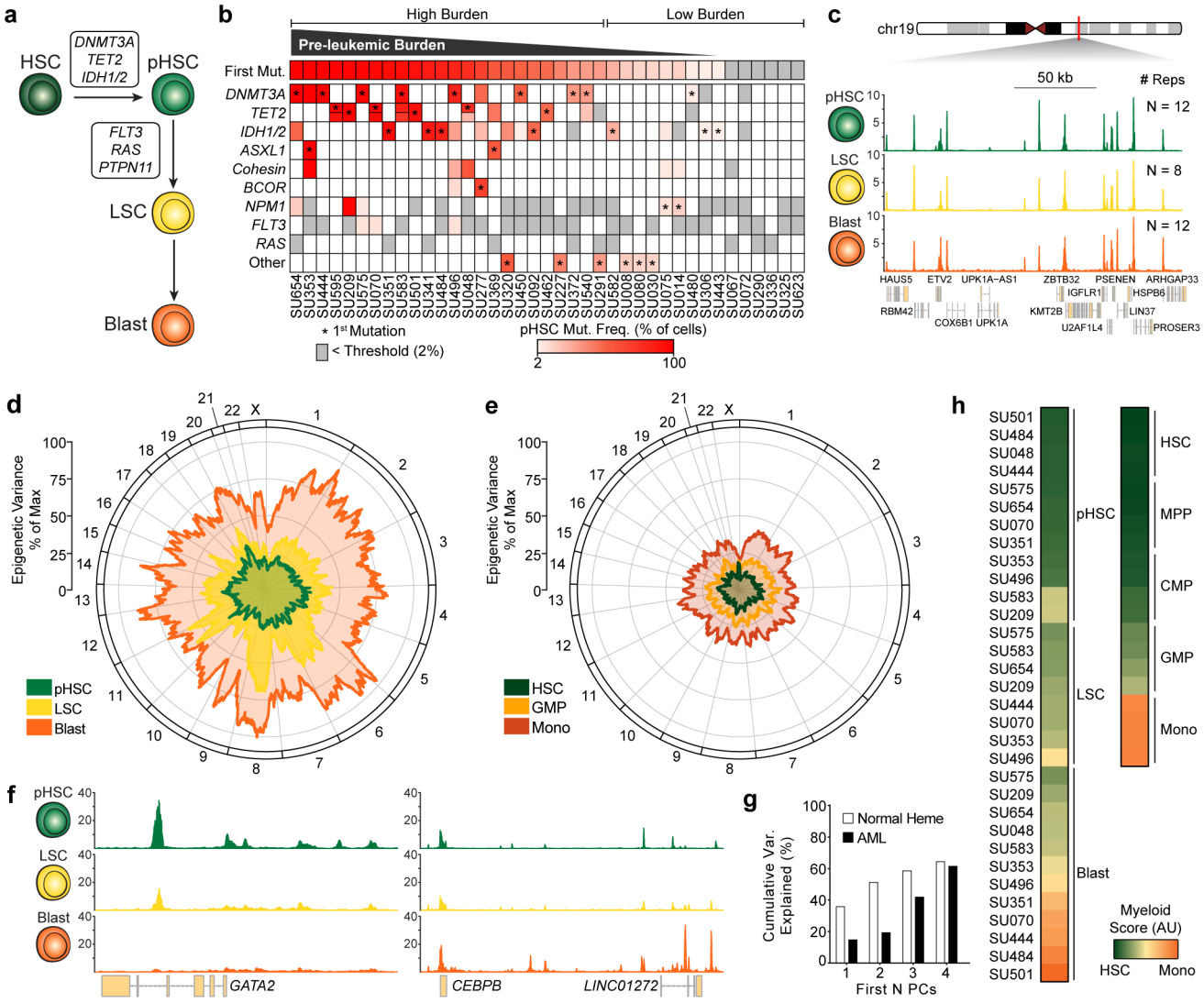

Figure 5. Acute myeloid leukemia regulomes reveal a cooption of normal myelopoiesis

(a) Schematic of the leukemogenic process.

(b) Mutation frequencies of HSCs isolated from AML patients (N=39). Color indicates the percent of cells harboring the given mutation as estimated from the variant allele frequency. Gray indicates a mutation present in leukemic cells but not observed in pHSC (i.e. a late mutation event, detection threshold = 2% of cells or 1% of alleles). Asterisks indicate the predicted first mutation. If a mutation is bi-allelic, the representative bar is divided in half.

(c) Normalized ATAC-seq profiles at a control locus (chr19:36102236–36277236) from FACS-purified AML cell types. Profiles represent the union of all biological replicates for each cell type. Y axis scale ranges from 0–10 in normalized arbitrary units.

(d,e) Mean variance of ATAC-seq signal across the linear genome as calculated by a moving average for (d) each leukemic cell stage and (e) their corresponding normal cell types (see methods). The distance from the center of the graph represents the variance. The position along the circumference represents the genomic position.

(f) Normalized ATAC-seq profiles near *GATA2* (left; chr3:128,197,777–128,218,433) and *CEBPB* (right; chr20:48,800,260–48,904,715). Profiles shown are for SU444. Y axis scale ranges from 0–10 in normalized arbitrary units.

(g) Cumulative variance of AML ATAC-seq data explained by the first N principal components derived from normal hematopoiesis.

(h) Myeloid development score derived from ATAC-seq data in normal blood cell types (N=4 biological replicates) and AML cell types.

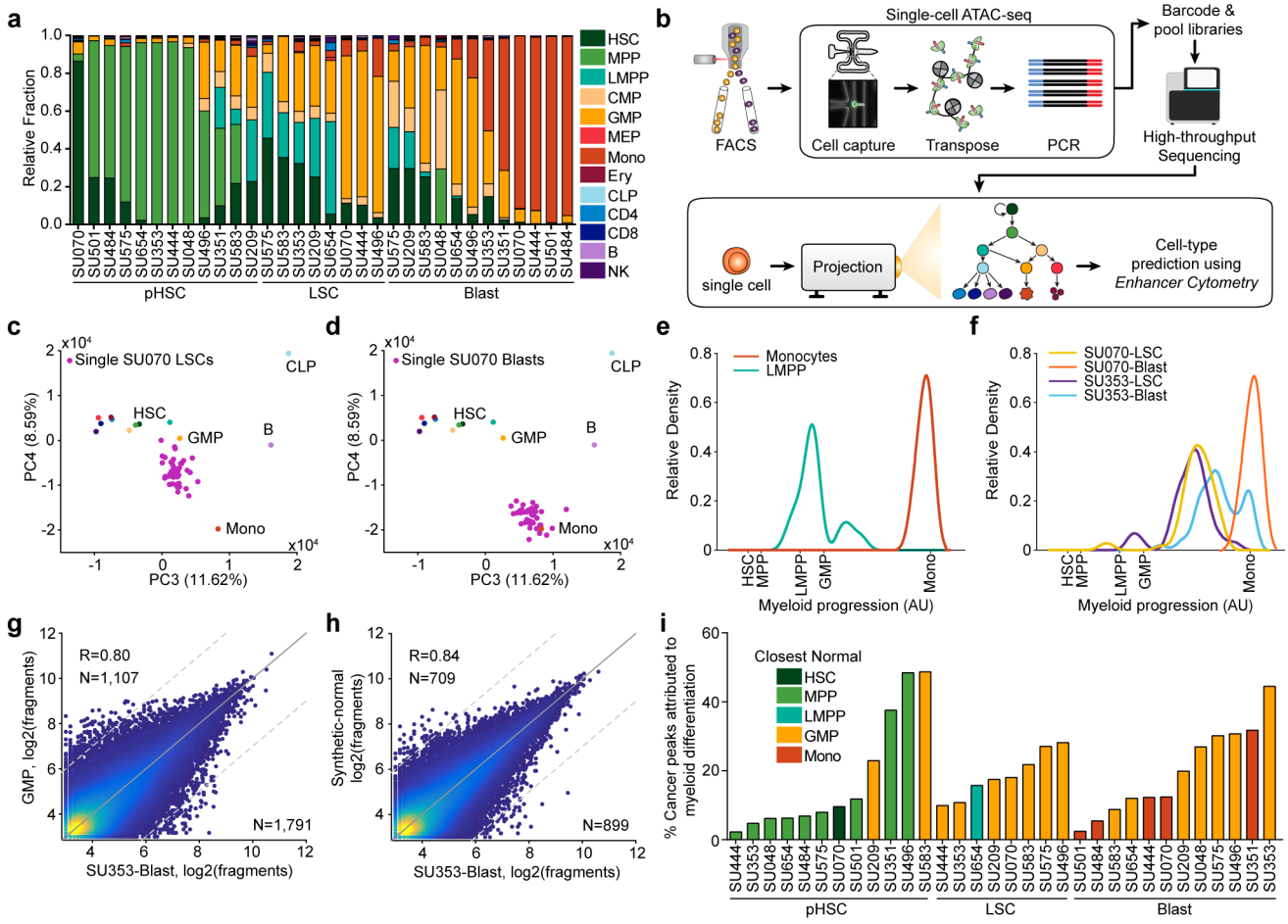

Figure 6. Enhancer cytometry and single-cell regulomes support a model of regulatory heterogeneity and allow for deconvolution of AML-specific biology

(a) Enhancer cytometry deconvolution showing the predicted contribution of various normal cell types to the regulatory landscape of different AML cell types.

(b) Schematic of single-cell ATAC-seq protocol and analysis (see methods).

(c,d) Projection of ATAC-seq data derived from (c) single SU070 LSCs (N=71) and (d) single SU070 blast cells (N=42) onto the principal components derived from the normal hematopoietic hierarchy.

(e,f) Relative density of (e) single LMPPs (N=68) and monocytes (N=90) and (f) single SU070 LSCs (N=62), SU070 blasts (N=42), SU353 LSCs (N=36), and SU353 blasts (N=52) projected onto a one-dimensional representation of the myeloid developmental progression.

(g,h) Scatter plot showing the correlation of ATAC-seq data derived from SU353 blast cells with (g) the closest normal cell type (GMP) (R=0.86) and (h) the enhancer cytometry-defined synthetic normal (R=0.91). Cutoff for differential peaks is a log2(fold change) greater than 3. The R values reported are calculated from correlations of all peaks. Plots show 50,000 random peaks, each with at least 5 reads.

(i) Comparison of AML cell types to synthetic normal analogs. For each sample, the closest normal cell type is indicated by the color of the bar. The percent of the total significant peaks

(called by closest normal comparison) that are removed by comparison to synthetic normal analogs is plotted for each sample.

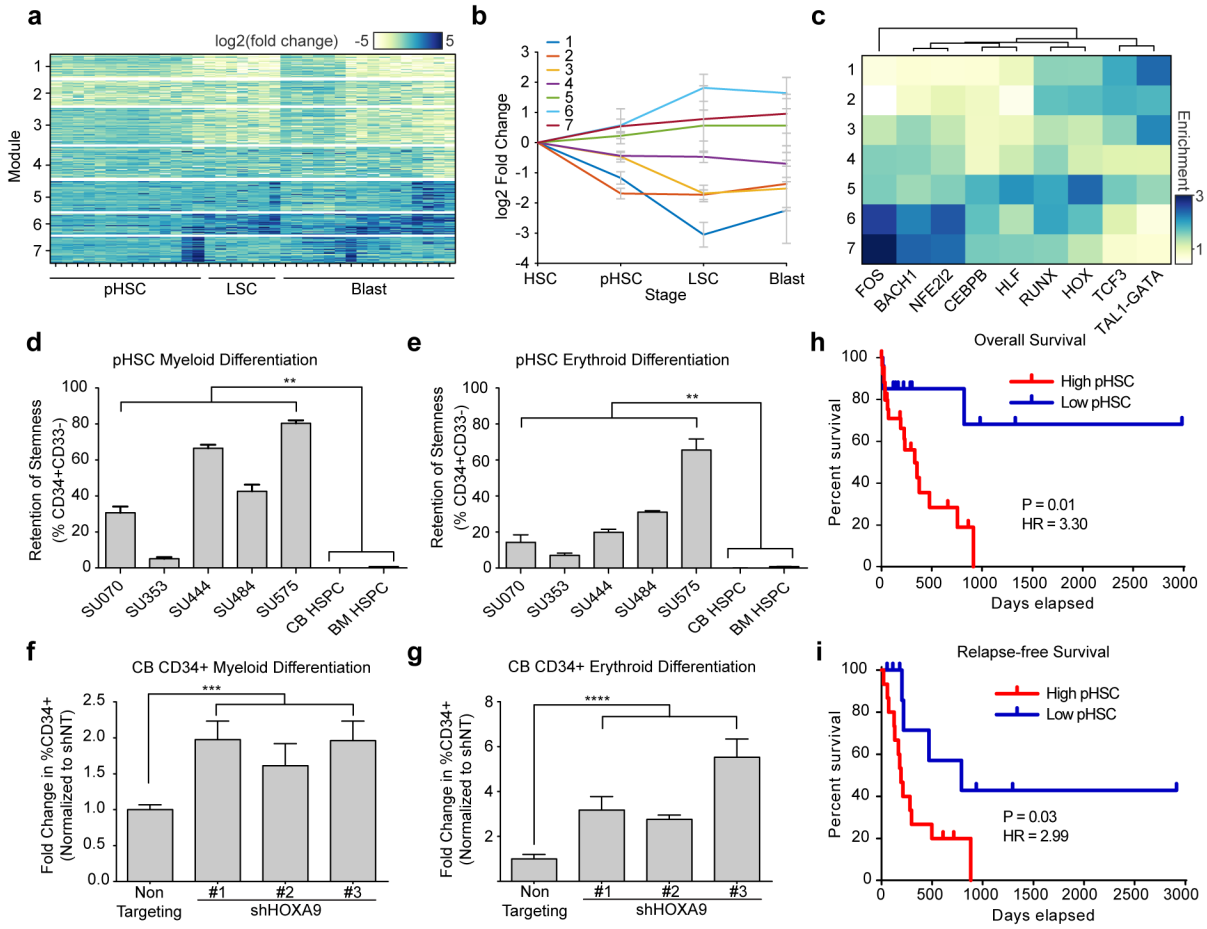

Figure 7. Early chromatin accessibility alterations within pHSCs cause defects in differentiation which correlate with adverse patient outcomes

(a) K-means clustering was used to identify 7 clusters of co-varying peaks, termed regulatory modules (see methods).

(b) Enrichment of each regulatory module shown in (a) at each stage of leukemia evolution. All biological replicates for each AML cell type were merged. Error bars shown represent 1 S.D. across all samples of that given cell type.

(c) Enrichment and hierarchical clustering of motifs enriched in each of the 7 AML-specific regulatory modules.

(d,e) Retention of CD34 expression as measured by flow cytometric analysis after 6 days of enforced differentiation down the (d) myeloid lineage and (e) erythroid lineage. Error bars represent 1 S.D. Experiments done in triplicate.

(f,g) Fold change in the percent of cells expressing CD34 as measured by flow cytometric analysis of cord blood-derived HSCs transduced with shRNAs targeting HOXA9 or a non-targeting control. CD34 expression was measured after 6 days of differentiation down the (f) myeloid or (g) erythroid lineage. Only GFP+ transduced cells analyzed. Error bars represent 1 S.D. Experiments done in triplicate.

(h) Overall and (i) relapse-free survival of patients stratified by pre-leukemic burden (High burden, N=24; Low burden N=15). High pre-leukemic burden defined as > 20% of HSCs

harboring at least the first pre-leukemic mutation. P values comparing two Kaplan-Meier survival curves were calculated using the log-rank (Mantel-Cox) test. Hazard ratios (HR) were determined using the Mantel-Haenszel approach.
**p<0.01, ***p<0.001, ****p<0.0001 derived from two-tailed t-test